DNA
플러스
2019

공공기관 신뢰가능 AI 구현 실용가이드

- OECD 권고안의 적용 -

작성 | 김동현 수석(정책본부 미래전략센터) kimdh@nia.or.kr
장준희 선임(정책본부 미래전략센터) junhee@nia.or.kr

자문 | 고상원 실장(정보통신정책연구원)
김명주 교수(서울여자대학교)
김연규 팀장(SK Telecom)
박준호 전문위원(LG전자)
이성웅 상무(IBM Korea)
이태희 상무(Microsoft Korea)
조원서 본부장(한국산업기술시험원)

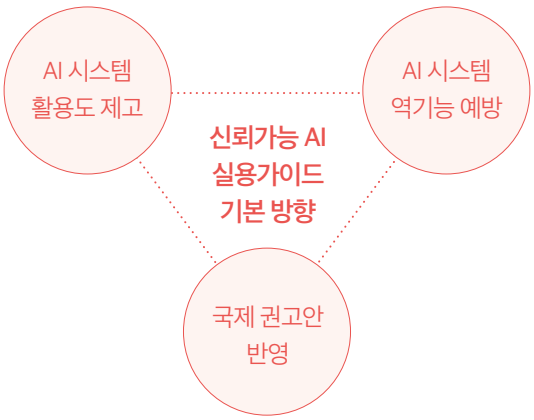
NIA 한국정보화진흥원
NATIONAL INFORMATION SOCIETY AGENCY

Contents

요약

공공기관 신뢰가능 AI의 구현 실행가이드 필요성

- 공공기관은 공공서비스 및 업무개선을 혁신할 AI 도입을 추진하고 있는 가운데 알고리즘·데이터 편향, 프라이버시 침해 등의 역기능도 증가할 것으로 우려
- 공공기관이 AI의 부정적 위험을 최소화하는 신뢰가능 AI를 구현할 수 있도록 돕는 실용가이드를 개발하고 적용 필요
- 아울러 OECD, EU 등 신뢰가능 AI에 관한 국제적 권고사항을 실용가이드에 반영, 국제 표준 요구사항 대응 및 준수 필요



[그림 1] 신뢰가능 AI 실용가이드 기본방향

신뢰가능한 AI 구현을 위한 원칙(OECD)

- OECD는 2019년 5월 'OECD AI 권고안'을 회원국 만장일치로 공식 채택, 신뢰가능 AI 구현을 위한 5가지 원칙과 신뢰가능 AI 시스템을 정의
  - ❶ 포용 성장, 지속가능 발전, 복지 증진
  - ❷ 인간중심 가치 지향, 공정성 지향
  - ❸ 투명성 확보, 설명 가능성 확보
  - ❹ 보안 및 안전성 확보
  - ❺ 책임성 확보

『신뢰가능 AI 시스템』이란 AI 시스템의 기획, 개발 단계부터 구축과 운영하는 단계까지 생태계 전반에 걸쳐 신뢰가능 AI의 제 원칙이 실현된 AI 시스템

공공기관 신뢰가능 AI의 구현 실행가이드

• 공공기관이 신뢰가능 AI 구현을 위한 원칙을 준수하기 위한 실행가이드

원칙	실행가이드
포용성장, 지속가능 발전, 복지증진	공공성의 확인 - AI 시스템의 기관 미션 연계성과 사회경제적 영향 평가
	사회적 차별요소 배제 - 데이터, 모델로부터 성, 인종 등 차이로 인한 근원적 차별 배제
인간중심 공정성	인간중심 가치와 공정성 촉진 - 인권영향평가, 인권실사, 윤리 행동 강령, 품질인증 조치
	인간중심 가치 내재화 - 적절한 안전장치 (Kill Switch, Human in the loop 등)
투명성 설명가능성	AI 시스템에 대한 투명한 정보공개 AI에 관한 일반 정보, 개발/훈련/ 운영/활용의 방식에 관한 정보
	AI 시스템 결과에 대한 설명 요인, 데이터, 알고리즘 등 의사결정 요인과 전후 맥락 설명
보안 및 안전성	AI 시스템의 추적 가능성 보장 - 데이터 세트, 알고리즘, 프로세스 및 의사결정 관련 추적 가능성
	체계적인 위험관리 접근 - 가능한 위험 및 확률, 관리방안
책임성	AI 시스템 원칙의 실현 - 라이프사이클에서 발생한 의사결정과 행동 문서화

신뢰가능한 AI 구현 거버넌스 프레임워크

• 공공기관이 신뢰가능 AI 구현 실행에 관한 수행방법 가이드

방법	세부 지침
AI 거버넌스 구축과 운영	신뢰가능 AI 시스템 구현을 위한 협력과 소통 유도
	AI 시스템의 위험 평가 및 내부 관리
AI 시스템 위험평가와 의사결정 모델	AI 시스템의 위험의 확률과 심각성 평가
	AI 시스템 의사결정 프로세스에 인간의 개입 수준 결정
책임성 있고 우수한 데이터관리	데이터 이력관리
	데이터 품질관리
	데이터에 내재하고 있는 편향의 최소화
소비자 및 고객과 커뮤니케이션	AI 시스템 활용 관련 투명한 정보공개
	사람-AI 인터페이스 정보
	설명 정책 수립 및 추진

1 실용가이드 개요

1 배경 및 목적

공공기관 AI 도입 확대 전망

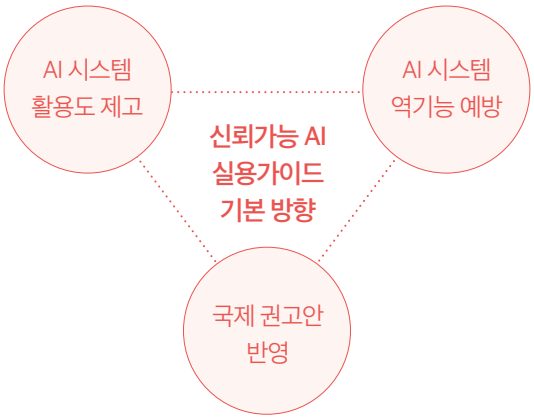
- 공공기관은 폭발적으로 증가하는 가용정보를 분석하고 기관의 의사결정과 공공서비스를 혁신할 AI 도입을 적극 추진하고 있음
- 이미 미국 등 AI 선진국은 AI를 활용하여 건강, 교통, 환경, 치안의 다양한 분야에서 혁신적 서비스와 혜택을 시민에게 제공하고 있음

AI 역기능 증가 우려 및 대응 필요

- AI가 확산되면서 알고리즘·데이터 편향에 의한 편견, 프라이버시 침해 등의 역기능도 증가할 것으로 우려
- 공공기관은 AI 역기능이 국민에게 직접적으로 영향을 미칠 수 있다는 점에서 다양한 이슈를 선제 연구하고 대응 필요

신뢰가능 AI 구현을 위한 실용가이드 개발 필요

- 공공기관이 AI의 긍정적 효과를 극대화하고 부정적 위험을 최소화하는 신뢰가능 AI를 구현할 수 있도록 돕는 실용가이드를 개발하고 적용 필요
- 아울러 OECD, EU 등 신뢰가능 AI에 관한 국제적 권고사항을 실용가이드에 반영, 국제 표준 요구사항 대응 및 준수 필요



[그림 1] 신뢰가능 AI 실용가이드 기본방향

2 구성

- 공공기관의 신뢰가능 AI 구현을 위한 실용가이드에서 소개하는 주요 내용은 다음과 같음
  - 신뢰가능 AI의 개념과 기본원칙
  - 기본원칙 준수를 위한 고려사항(considerations)과 실행가이드
  - 신뢰가능 AI 구현을 위한 관리방법(measures)과 거버넌스 프레임워크

3 활용

- 본 실용가이드는 모든 공공기관이 적용가능한 범용 가이드로 작성
- 공공기관이 신뢰가능 AI 구현을 목적으로 본 실용가이드를 활용하는 경우, 기관의 특성에 맞게 맞춤 제작하여 활용할 수 있음
  - 이 경우 본 실용가이드를 인용하였음을 표시하여야 함
- 본 실용가이드는 AI 기술 발전속도와 파급효과에 맞춰 민첩하게 그리고 지속적으로 개정할 계획임

2 신뢰가능 AI 기본원칙

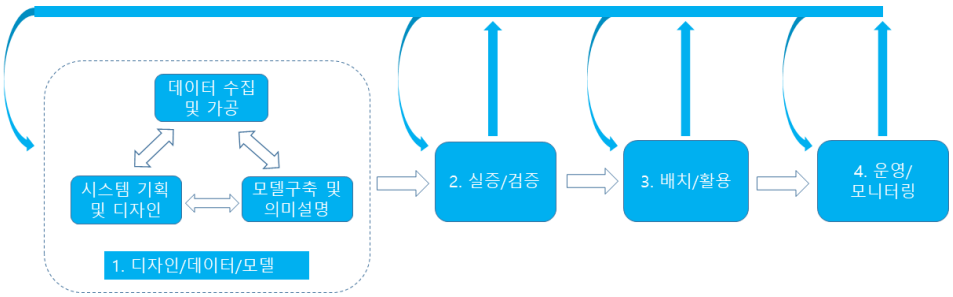
1 AI 관련 주요 용어

• 『AI 시스템』은 인공지능 기술을 활용하여 실상 또는 가상 환경에 영향을 미치는 예측과 의사결정을 내릴 수 있는 시스템

※ AI 시스템은 기존 정보시스템과 달리 인간이 정의한 목표에 따라 다양한 수준의 자율성을 발휘하도록 설계될 수 있음

• 『AI 시스템 라이프사이클』은 크게 (ⅰ)디자인/데이터/모델, (ⅱ)실증/검증, (ⅲ)배치/활용, (ⅳ)운영/모니터링 단계로 구성

※ 라이프사이클의 단계는 AI 특성상 순차적인 것은 아닐 수 있으며 중간 단계에서도 그 이전 단계로 반복적으로 환원될 수 있음



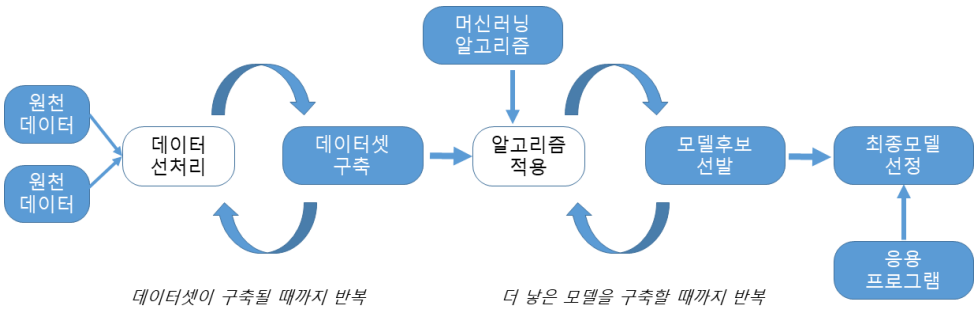
[그림 2] AI 시스템 라이프사이클의 기본단계

- 디자인/데이터/모델 단계는 다음 세부단계로 나눌 수 있음

· AI 시스템 기획과 디자인 : AI 시스템의 컨셉, 목적, 전제조건, 내용, 필수조건, 시스템 프로토타입에 관한 기획과 디자인

· 데이터 수집과 가공 : 데이터의 수집, 가공, 데이터 품질 검사, 데이터셋 특성과 메타데이터(출처, 특성, 관리)의 문서화

· 모델 선택과 의미설명 : 모델(알고리즘)의 생성과 선택, 교정, 교육 및 모델에 대한 의미 설명



[그림 3] AI 데이터 디자인/데이터/모델 단계

- 실증/검증 단계는 다양한 차원과 환경에서 모델 성능 및 특성 평가, 테스트를 통한 모델 실증 및 결과에 따른 모델 조정 포함
- 배치/활용 단계는 파일럿 적용, 레거시 시스템과 결과 일치성 확인, 규정 준수 여부, 조직 변화 관리 및 사용자 경험 평가 포함
- 운영/모니터링 단계는 AI 시스템 목표와 AI 윤리의 관점에서 AI 시스템의 운영평가, 결과영향평가 및 지속적 업데이트 포함

※ 각 단계에서 심각한 문제가 발견된 경우 해결을 위해 이전 단계로 환원하거나 필요한 경우 AI 시스템을 폐기할 수 있음

• 『AI 행위자』는 앞의 AI 시스템 라이프사이클의 각 단계에서 적극적인 역할을 수행하는 행위주체인 사람 또는 기관

• 『AI 이해관계자』는 AI 행위자를 포함하여 AI 시스템과 연관되어 있거나 영향을 받는 모든 사람 또는 기관
  - AI 시스템 사용자는 기관의 내부업무처리와 외부고객 서비스를 위해 AI 시스템을 사용하는 기관의 직원을 의미함
  - AI 시스템 소비자는 AI 시스템이나 AI 기능이 탑재된 서비스 또는 제품을 사용함으로 영향을 받는 소비자임과 동시에
  - AI로 강화된 서비스를 받기 위하여 AI의 모델과 데이터분석에 필요한 개인정보를 제공하는 고객을 의미함



1. AI 시스템 행위자
- AI 응용시스템 개발/구축/운영
- AI 기능이 탑재된 공공서비스의 개발

2. AI 시스템 사용자(기관)
- AI 시스템을 내부 업무 프로세스에 활용
- AI 시스템을 활용한 공공서비스 제공

3. 소비자 및 고객
- AI 시스템이나 AI 서비스의 소비자 및 고객
- 강화된 AI 모델개발을 위한 개인정보 제공자

[그림 4] AI 시스템 가치사슬과 이해관계자

• 『AI 지식』은 AI 시스템 라이프사이클을 이해하는데 필요한 데이터, 코드, 알고리즘, 모델, 연구, 노하우, 훈련프로그램, 거버넌스, 프로세스 및 모범 사례와 같은 기술과 리소스를 의미

2 OECD의 신뢰가능 AI의 원칙

• OECD는 'OECD AI 권고안(2019.5)'을 통해 다음과 같이 신뢰가능 AI 구현을 위한 5가지 원칙과 신뢰가능 AI 시스템을 정의

원칙 1. 포용 성장, 지속가능 발전과 복지 증진

• 모든 AI 이해관계자는 인류의 포용 성장, 지속 가능 발전 및 복지 증진을 위해 신뢰가능한 AI의 구현에 힘써야 함

- AI 이해관계자는 인간의 능력과 창의력을 향상시키고 소수집단 포용을 진전시키는 방향으로 AI의 구현에 힘써야 함
- 아울러 성차별 등 사회적 불평등을 감소시키고, 자연 환경을 보호하는 방향으로 AI의 구현에 힘써야 함

원칙 2. 인간중심 가치와 공정성

• AI 행위자는 AI 시스템 라이프사이클 전반에 걸쳐 법률, 인권, 민주적 가치, 공정 등 인간 중심 가치를 존중하고 지켜나가야 함

- AI 행위자는 자유, 존엄, 자치, 사생활 보호, 평등, 다양성, 공정성, 차별 금지, 노동권보장의 인간중심 가치를 지켜나가야 하며
- 이를 위해 AI 행위자는 인간중심 가치 실현을 위한 메카니즘과 인간이 최종의사결정에 개입할 수 있는 안전장치를 마련해야 함

원칙 3. 투명성과 설명가능성

- AI 행위자는 사용자와 고객에게 AI 시스템에 대한 의미 있는, 최신 정보를 제공함으로써 그들과 적극적으로 의사소통해야 함
  - AI 행위자는 AI 시스템 활용 시점, AI 시스템 개발, 배치 및 운영 방식에 대해 투명하게 정보를 제공할 수 있어야 함
  - AI 행위자는 이해관계자에게 AI 시스템의 예측, 의사결정의 기저가 되는 핵심요인과 논리에 대하여 쉽게 설명할 수 있어야 함

원칙 4. 보안과 안전성

- AI 시스템은 전 수명주기에 걸쳐 어떠한 조건에서도 견고하게 작동되어야 하며 외부에 취약점이 노출되지 않아야 함
  - AI 행위자는 AI 시스템의 결과와 반응을 분석하기 위하여 데이터셋, 프로세스, 의사결정 과정 등을 추적할 수 있어야 함
  - AI 행위자는 개인정보보호, 정보보안, 외부공격의 위험을 해소하기 위하여 체계적으로 위험관리방법을 지속적으로 적용하여야 함

원칙 5. 책임성

- AI 행위자는 책임지고 AI 시스템 구현에 있어 위의 원칙을 실현하는 것과 AI 시스템이 올바르게 기능할 수 있도록 노력해야 함
  - 신뢰가능 AI 시스템 구현에 있어서 AI 행위자가 윤리적, 도덕적으로 행동할 수 있도록 안내하고 행동강령 등을 명시함
  - AI 행위자는 관련 문서를 제공하거나 경우에 따라서 감사를 받음으로 자신의 책임성에 입증할 수 있어야 함

3 신뢰가능 AI 시스템 정의

『신뢰가능 AI 시스템』이란 AI 시스템의 기획, 개발 단계부터 구축과 운영하는 단계까지 생태계 전반에 걸쳐 신뢰가능 AI 원칙의 가치가 고려되고 실현된 AI 시스템 (OECD AI 권고안)

3 공공기관 신뢰가능 AI의 구현 실행가이드

실행가이드 요약

AI 원칙 실행가이드 \ AI 활동		기획 기획자	디자인 설계자	데이터 관리자	모델 개발자	검사실증 기획자/감리사	도입/설치/운영 운영자
포용성장, 지속가능 발전, 복지증진	공공성의 확인 - AI 시스템의 기관 미션 연계성과 사회경제적 영향 평가	✓				✓	
	사회적 차별요소 배제 - 데이터, 모델로부터 성, 인종 등 차이로 인한 근원적 차별 배제		✓	✓	✓		✓
인간중심 공정성	인간중심 가치와 공정성 촉진 - 인권영향평가, 인권실사, 윤리 행동 강령, 품질인증 조치	✓				✓	✓
	인간중심 가치 내재화 - 적절한 안전장치 (Kill Switch, Human in the loop 등)		✓	✓	✓		✓
투명성 설명가능성	AI 시스템에 대한 투명한 정보공개 AI에 관한 일반 정보, 개발/훈련/ 운영/ 활용의 방식에 관한 정보		✓	✓	✓		✓
	AI 시스템 결과에 대한 설명 - 요인, 데이터, 알고리즘 등 의사결정 요인과 전후 맥락 설명		✓	✓	✓		✓
보안 및 안전성	AI 시스템의 추적 가능성 보장 - 데이터 세트, 알고리즘, 프로세스 및 의사결정 관련 추적 가능성		✓	✓	✓		✓
	체계적인 위험관리 접근 - 가능한 위험 및 확률, 관리방안	✓	✓	✓	✓	✓	✓
책임성	AI 시스템 원칙의 실현 - 라이프사이클에서 발생한 의사 결정과 행동 문서화	✓	✓	✓	✓	✓	✓

1 원칙 1. 포용 성장, 지속가능 발전과 복지 증진

• 모든 AI 이해관계자는 국내외적으로 인류의 포용 성장, 지속 가능 발전 및 복지 증진을 위해 신뢰가능 AI의 구현에 힘써야 함

AI 원칙 실행가이드 \ AI 활동		기획	디자인	데이터	모델	검사실증	도입/설치/운영
		기획자	설계자	관리자	개발자	기획자/감리사	운영자
포용성장, 지속가능 발전, 복지증진	공공성의 확인 - AI 시스템의 기관 미션 연계성과 사회경제적 영향 평가	✓				✓	
	사회적 차별요소 배제 - 데이터, 모델로부터 성, 인종 등 차이로 인한 근원적 차별 배제		✓	✓	✓		✓

AI 시스템의 공공성 확인

• AI 기획자/중간검수자는 AI 시스템 개발 전후로 다양한 이해관계자 협업과 토론을 통해 구현하려는 AI 시스템의 용도와 목적이 공공기관의 미션과 잘 부합하는지 확인해야 함

- 아울러 AI 시스템 구현을 통한 불평등, 환경오염, 인권차별 등의 부정적 영향이 일어날 가능성에 대한 논의도 함께 되어야 함

AI 시스템의 사회적 차별요소 배제

• AI 시스템의 설계자/관리자/개발자는 데이터와 AI모델로 부터 성, 인종, 지역 등의 차이로 인한 근원적 차별을 배제해야 함

- AI 시스템의 운영자는 AI 시스템 활용결과, 사회적 차별요소가 감지되었을 때, 즉시 AI 행위자에게 즉시 관련된 모든 정보를 제공하여야 함

2 원칙 2. 인간중심 가치와 공정성

• AI 행위자는 AI 시스템 라이프사이클 전반에 걸쳐 법률, 인권, 민주적 가치, 공정 등 인간 중심 가치를 존중하고 지켜나가야 함

AI 원칙 실행가이드 \ AI 활동		기획	디자인	데이터	모델	검사실증	도입/설치/운영
		기획자	설계자	관리자	개발자	기획자/감리사	운영자
인간중심 공정성	인간중심 가치와 공정성 촉진 - 인권영향평가, 인권실사, 윤리 행동강령, 품질인증 조치	✓				✓	✓
	인간중심 가치 내재화 - 적절한 안전장치 (Kill Switch, Human in the loop 등)		✓	✓	✓		✓

AI 시스템의 인간중심 가치와 공정성 촉진

• AI 시스템 기획자/중간검수자/운영자는 윤리적 행동강령을 마련하는 등 인간중심 가치와 공정성 촉진에 대한 약속을 분명히 하여야 함

- 아울러 상황에 따라 인권영향평가(참고 1) 및 인권실사(참고 2), 등의 조치를 취하여 잠재적 위험의 심각성, 인권에 미치는 영향을 식별해야 함

AI 시스템에 인간중심 가치 내재화

• AI 시스템의 설계자/관리자/개발자는 인간중심 가치와 공정성을 AI 시스템에 내재화하여 개발하는 것이 중요함

- 아울러 AI 시스템의 설계자/관리자/개발자는 상황에 따라 사람이 개입(human-in-the-loop)할 수 있고 감독할 수 있게 하는 등 적절한 안전장치를 갖춘 AI 시스템을 개발하여야 함(참고 3)

3 원칙 3. 투명성과 설명가능성

• AI 행위자는 사용자와 고객에게 AI 시스템에 대한 의미 있는, 최신 정보를 제공함으로써 그들과 적극적으로 의사소통해야 함

AI 원칙 실행가이드 \ AI 활동		기획	디자인	데이터	모델	검사실증	도입/설치/운영
		기획자	설계자	관리자	개발자	기획자/감리사	운영자
투명성 설명가능성	AI 시스템에 대한 투명한 정보공개 AI에 관한 일반 정보, 개발/훈련/ 운영/ 활용의 방식에 관한 정보		✓	✓	✓		✓
	AI 시스템 결과에 대한 설명 - 요인, 데이터, 알고리즘 등 의사결정 요인과 전후 맥락 설명		✓	✓	✓		✓

AI 시스템에 대한 투명한 정보 공개

• AI 시스템 설계자/개발자/운영자는 AI 시스템이 어떻게 개발되고, 훈련되며, 운영 및 활용되는지 방식을 이해관계자에게 공개

- 그러나 소스코드 및 데이터 세트는 영업비밀 또는 개인정보보호 등 지적 재산권의 보호 적용을 받을 수 있음

AI 시스템 결과에 대한 설명

• AI 시스템 설계자/개발자/운영자는 AI 시스템이 의사결정에 어떻게 도달했는지에 관하여 이해관계자에게 설명할 필요가 있음

- 의사 결정의 주요 변수, 결정 요인, 데이터, 논리 또는 알고리즘을 명확하고 간단한 용어로 문맥에 따라 적절하게 제공
- 아울러 AI 시스템 특성상 유사한 환경임에도 불구하고 다른 결과를 생성한 이유를 설명할 필요가 있음

4 원칙 4. 보안과 안전성

• AI 시스템은 전 수명주기에 걸쳐 어떠한 조건에서도 견고하게 작동되어야 하며 외부에 취약점이 노출되지 않아야 함

AI 원칙 실행가이드 \ AI 활동		기획	디자인	데이터	모델	검사실증	도입/설치/운영
		기획자	설계자	관리자	개발자	기획자/감리사	운영자
보안 및 안전성	AI 시스템의 추적 가능성 보장 - 데이터 세트, 알고리즘, 프로세스 및 의사결정 관련 추적 가능성		✓	✓	✓		✓
	체계적인 위험관리 접근 - 가능한 위험 및 확률, 관리방안	✓	✓	✓	✓	✓	✓

AI 시스템 추적 가능성 보장

• AI 시스템 설계자/개발자/운영자는 AI 시스템 라이프사이클 동안 이루어진 데이터 세트, 프로세스 및 의사 결정과 관련하여 추적이 가능하다는 것을 보장해야 함

- AI 시스템의 결과 및 문의에 대한 응답을 상황에 적합하고 최신 기술과 일치하도록 분석해야 함

AI 시스템 체계적인 위험관리 접근

• AI 행위자 모두는 각자의 역할, 상황 및 행동 능력에 따라 AI 시스템 수명주기의 각 단계마다 체계적인 위험 관리 접근 방식을 지속적으로 적용해야 함

- 위험관리에서 다루어야 할 위험으로 AI 시스템 고유의 위험, 개인 정보 보호, 디지털 보안, 데이터 편견 등이 있음

5 원칙 5. 책임성

• AI 행위자는 책임지고 AI 시스템 구현에 있어 위의 원칙을 실현하는 것과 AI 시스템이 올바르게 기능할 수 있도록 노력해야 함

AI 원칙 실행가이드 \ AI 활동		기획	디자인	데이터	모델	검사실증	도입/설치/운영
		기획자	설계자	관리자	개발자	기획자/감리사	운영자
책임성	AI 시스템 원칙의 실현 - 라이프사이클에서 발생한 의사결정과 행동 및 조치 문서화	✓	✓	✓	✓	✓	✓

AI 시스템 원칙의 실현

• AI 행위자는 적용 가능한 윤리강령, 규정에 따라 앞서 언급한 신뢰가능 AI 구현을 위한 모든 원칙을 존중하고 그들의 행동과 의사 결정을 통해 이를 입증하여야 함

- 책임성 제고를 위해 AI 시스템 라이프사이클 동안 발생한 주요 결정 사항, 행동 및 조치를 문서화하여야 하고 기관내 윤리위원회, 규제위원회의 역할을 보장하고 적극 활용하여야 함

4 신뢰가능 AI 구현 거버넌스 프레임워크

거버넌스 프레임워크 요약

• 거버넌스 프레임워크는 앞 장에서 설명한 공공기관이 신뢰가능 AI 구현을 위한 조치를 어떻게 수행할 것인가에 대한 가이드

방법	세부 지침
AI 거버넌스 구축과 운영	신뢰가능 AI 시스템 구현을 위한 협력과 소통 유도
	AI 시스템의 위험 평가 및 내부 관리
AI 시스템 위험평가와 의사결정 모델	AI 시스템의 위험의 확률과 심각성 평가
	AI 시스템 의사결정 프로세스에 인간의 개입 수준 결정
책임성 있고 우수한 데이터관리	데이터 이력관리
	데이터 품질관리
	데이터에 내재하고 있는 편향의 최소화
소비자 및 고객과 커뮤니케이션	AI 시스템 활용 관련 투명한 정보공개
	사람-AI 인터페이스 정보
	설명 정책 수립 및 추진

1 AI 거버넌스 구축, 운영

• 공공기관은 신뢰가능 AI 구현을 위해 AI 행위자 사이의 협력과 자유로운 의사소통이 가능하게 하며
- 발생가능한 위험을 관리할 내부 AI 거버넌스를 구축, 운영할 것을 권장함
- 공공기관은 관련 전문 지식과 적절한 대표성을 갖춘, 다학제적 거버넌스 기구를 설립하는 것을 고려하고 아래의 직무를 수행하도록 지원함

신뢰가능 AI 시스템 구현을 위한 협력과 소통 유도

- 신뢰가능 AI 시스템 구현을 위한 원칙을 지켜나갈 수 있도록 AI 행위자 사이의 협력과 소통을 유도하고 적절한 조치를 취함
  - 공공기관 AI 시스템에 대한 소비자와 고객의 신뢰를 받기 위해 정보공개 및 피드백 채널 확보 등 고객과의 소통을 보장

AI 시스템의 위험 평가 및 내부 관리

- 공공기관의 자체 위험관리 방법을 사용하여 아래와 같이 위험을 파악하고 관리하며 필요에 따라 제어함
  - AI 시스템 활용을 통해 개인에게 미칠 수 있는 모든 부정적인 위험을 파악 : 영향규모, 취약계층, 영향을 받는 방식, 영향을 받는 사람들로부터 피드백을 받는 방법 등
  - 각 위험의 발생확률과 영향도(심각성)를 고려하여 AI 시스템의 의사결정 모델에 사람의 개입 정도와 관련 조치를 결정

2 AI 시스템 의사결정(인간의 개입) 모델 결정

- 공공기관은 AI 시스템을 구현하기에 앞서 AI 시스템 활용 목적 대비 활용에 따른 위험(확률, 크기)을 평가하여야 함
  - 위험 평가 결과를 기반으로 AI 시스템에 대한 인간의 감독 또는 의사결정 개입의 수준(개입하지 않음/지속적 모니터링 필요/사람이 의사결정해야 함 등)을 결정함(참고 3, 4)

AI 시스템의 위험의 확률과 심각성 평가

- 공공기관의 의사결정(AI 시스템의 결정 또는 개인의 결정)의 결과로 개인에게 피해를 끼칠 확률과 크기를 분류
  - 피해를 어떻게 정의하느냐에 따라, 상황에 따라 피해 확률과 크기의 계산은 다르게 나타날 수 있음

AI 시스템 의사결정 프로세스에 인간의 개입 수준 결정

- AI 시스템의 조직 의사결정 프로세스에서 위험에 따른 피해의 확률과 심각성을 고려하여 인간의 개입 수준을 결정
  - AI 시스템의 의사결정 프로세스에서 인간의 개입 수준을 '개입하지 않음(human-out-of-the-loop)', 감독해야 함(human-over-the-loop)', 인간이 결정함(human-in-the-loop)의 세가지로 구별할 수 있음
  - 피해확률이 높고 심각성이 높으면 최종의사결정을 인간이 내리는 것이 맞으며, 반대로 피해확률도 낮고 심각성도 낮으면 인간이 개입하지 않는 것이 바람직함

3 책임성 있고 우수한 데이터 관리

- AI 모델을 구축하는 데 사용되는 데이터세트는 여러 소스에서 제공되며 AI 솔루션의 성공을 위해서는 데이터의 품질이 중요
  - 편향되거나 부정확하거나 대표성이 부족한 데이터를 사용하여 모델을 구축하면 모델에서 의도하지 않은 차별적 결정의 위험이 높음

데이터 이력관리

- 데이터 이력관리를 통해 데이터 출처 및 데이터가 소스에서 대상으로 이동하는 경로, 변환되는 방식, 다른 데이터와 상호 작용하는 위치 및 표현이 어떻게 변경되는지 등을 추적
  - 외부에서 얻은 데이터인 경우 데이터의 출처를 설정하기 어려울 수 있으며 이러한 외부 데이터 사용의 위험을 평가하고 그에 따라 관리해야 함

데이터 품질관리

- 데이터 세트의 정확성(accuracy), 완전성(completeness), 신뢰성(veracity), 최신성(Recency), 무결성(integrity), 활용적합성(usability), 라벨링(labeling) 등 데이터 품질에 영향을 주는 요소 관리 필요

데이터에 내재하고 있는 편향의 최소화

- 공공기관은 AI 시스템에 제공하는 데이터가 편향되어 있는지 파악해야 하며 그러한 편견을 완화하기 위한 조치를 취해야함

- 편향의 종류는 크게 5가지이며 편향 데이터를 파악하는 것이 쉽지 않으나 각 종류마다 최소화할 수 있는 방안을 활용(참고 5)

4 소비자 및 고객과 커뮤니케이션

• 적절한 커뮤니케이션은 기관과 개인 사이의 열린 관계를 구축하고 유지하면서 상호 신뢰를 강화

- AI 시스템을 구현할 때 소비자 및 고객과 커뮤니케이션 전략을 효과적으로 구현하고 관리하기 위해 다음 요소를 고려해야 함

AI 시스템 활용 관련 투명한 정보공개

• AI를 도입, 활용할 때 기관과 사람 간의 신뢰를 더욱 강화하기 위한 소비자 및 고객과 원활한 의사소통 방안 마련 필요

- 공공기관은 AI가 제품/서비스에 사용되는지에 대한 일반적인 정보(의사결정 방법, 역할 및 범위 등)와 AI의사결정이 개인에게 어떤 영향을 미치는지 등 투명성 제고를 위한 정보공개 필요

사람-AI 인터페이스 정보

• 공공기관은 소비자가 사람이 아닌 AI와 인터페이스하고 있는 것을 소비자에게 알릴 필요가 있으며

- 소비자의 응답이나 회신이 AI 시스템의 훈련에 사용될 것임을 알려야 하며, 소비자가 의도적으로 불량한 응답을 하는 경우 AI시스템 의 학습에 악영향을 줄 수 있음을 알려야 함

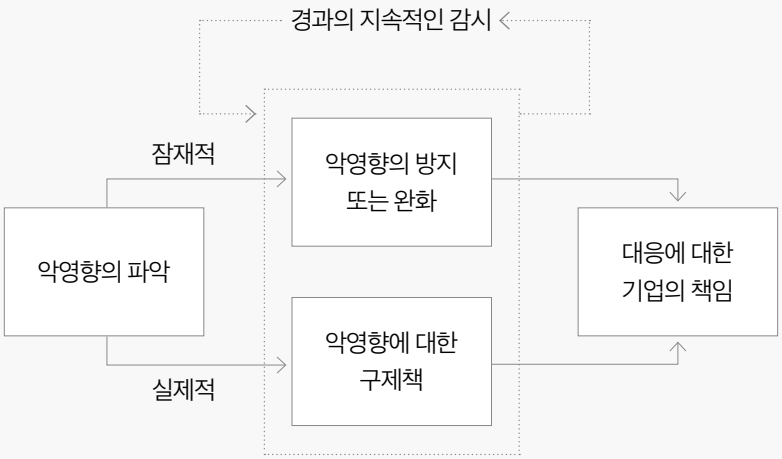
설명 정책 수립 및 추진

• 공공기관은 의사결정 프로세스에서 AI의 작동방식, 특정 결정이 내려진 방법 및 결정의 이유, 결정의 영향 및 결과에 대한 설명 필요

- 특히 소비자 또는 고객이 특정 결정에 관해 설명을 요청할 수 있는 제도적 장치를 마련하는 것이 필요함

[붙임]

[참고 1] 인권 실사(human rights due diligence)의 절차



[그림 설명]

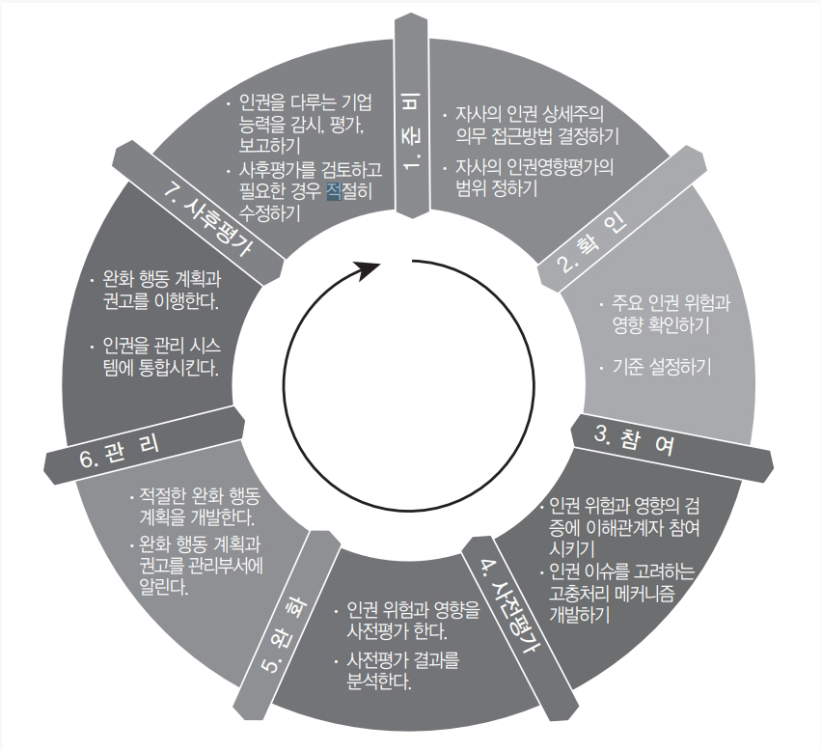
- 첫째 단계인 인권에 대한 악영향의 파악은 기관이 기관 내부 또는 기관 외부에서 자신의 기관 활동과 관련하여 인권 침해가 이루어지고 있는지 여부를 조사하는 것으로 일회적인 것에 그쳐서는 안 되고, 조사의 신뢰성을 높이기 위해 외부 전문가를 활용하는 것이 바람직함

- 둘째 단계는 인권 침해가 파악이 되면 그것을 방지 또는 완화하거나 그에 대한 구제책을 마련하여 인권 침해에 대한 대응을 하는 단계로 인권에 대한 악영향은 실제적으로 일어나는 것일 수도 있고, 단지 잠재적인 것일 수도 있는데, 기관은 각각의 경우에 맞는 대응책을 마련해야 함

- 셋째 단계에서 기관은 인권 침해 상황에 대한 자신의 대응을 지속적으로 감시하는데, 이러한 감시가 필요한 이유는 기관의 대응방식이 적절하였는지를 평가하기 위해서 뿐만 아니라, 인권 침해 상황이 재발되는 것을 막기 위함

- 마지막 넷째 단계에서는 기관의 사회적 책임을 실현하는 방법으로 기관이 파악하고 있는 인권 실태를 대외적으로 공표하고, 특히 심각한 인권 침해 상황에 관해서는 정기적으로 보고하도록 하는 단계로 기관이 인권 보호와 관련해서 외부와 소통하는 것을 보장하기 위해 이해관계자와의 협의도 이 단계에서 필요함

[참고 2] 인권영향평가(HRIAs) 과정



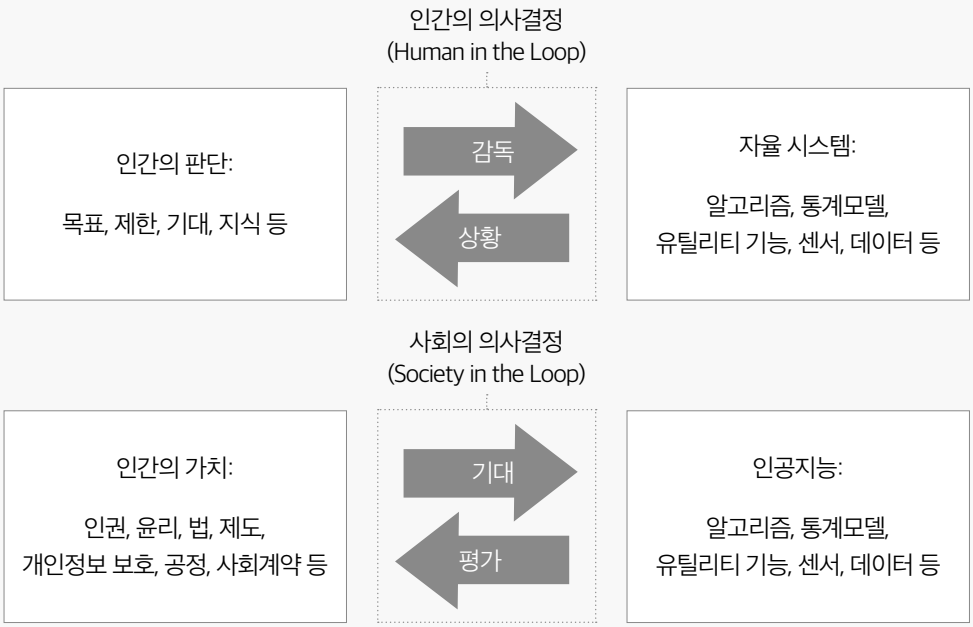
[그림 설명]

- 인권영향평가 과정은 대체로 계획 - 실행 - 평가 - 채택 과정을 거치게 된다. 하지만 대상과 상황마다 구체적인 평가 과정은 다를 수 밖에 없다.
  - International Finance Corporation(IFC)가 International Business Leaders Forum (IBLF)와 유엔 글로벌 콤팩트(UN Global Compact)의 협조 하에 작성한 '인권영향평가 및 관리 지침서 (HRIAM: Guide t

- Human Rights Impact Assessment and Management)' 에서는 인권영향평가를 1.준비 2.확인 3.개입 4.평가 5.개선 6.운영 7.최종평가로 이어지는 7 단계의 과정으로 수행할 것을 권고

- 여기서 중요한 것은 인권영향평가는 결코 일회성으로 끝날 수는 없다는 것이며 국내의 다른 영향평가, 대표적으로 환경영향평가나 교통영향평가가 사업 시작 전에 일회 실시하는 것으로 운영되는데 반해 인권영향평가는 일회성 평가를 탈피, 지속적으로 수행하는 것이 바람직함

[참고 3] 인간의 의사결정과 사회의 의사결정

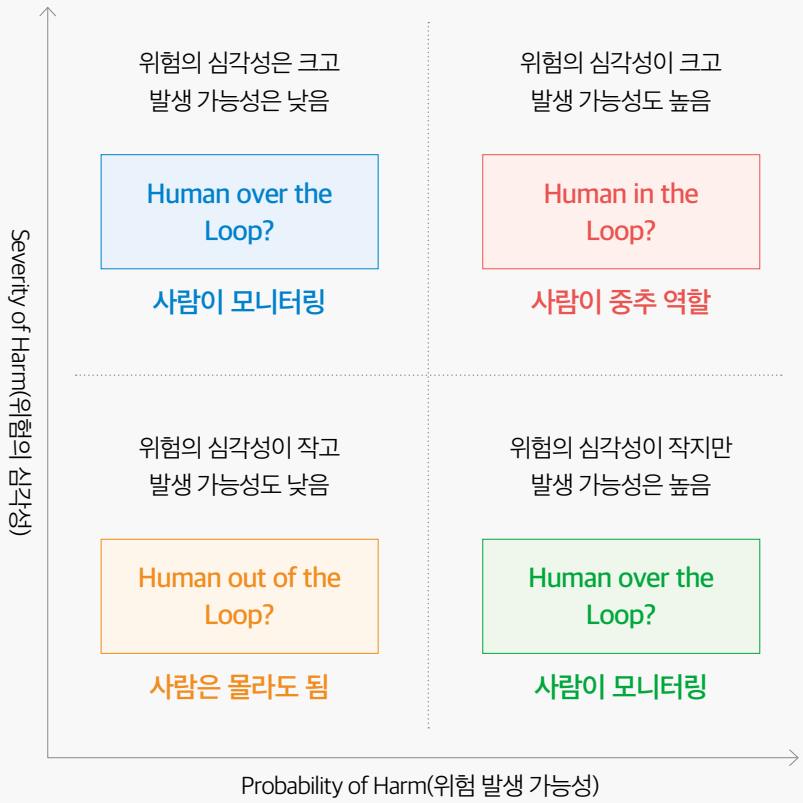


[그림 설명]

- '루프 안의 인간'은 알고리즘 개발 기관에 소속된 인간의 감독 하에 머신러닝 알고리즘을 실행하는 방식으로 이 원리의 절차는 다음과 같음
  - 개발한 머신러닝 알고리즘이 얼마나 정확하고 올바른 판단을 내렸는지에 대한 신뢰점수를 매기고, 신뢰점수가 기준보다 낮으면 알고리즘 데이터는 감독자에게 보내짐
  - 감독자는 신뢰도 평가 결과 리뷰 및 필요한 조치를 한 후 다시 테스트 과정을 반복하여 알고리즘을 완성
  - 하지만 이 경우 알고리즘 기준을 정하고 평가하는 감독자는 특정인으로 제한되는데, 이 때문에 소수 개발자의 사회적 편견이 알고리즘에 반영되는 문제가 발생할 수 있음. 이러한 잠재적 문제점 해결을 위해 알고리즘 개발자의 다양성을 확보하는 것이 과제

- MIT Lab 소속의 라완(Rahwan, 2016)은 '루프 안의 인간'의 대안으로 '루프 안의 사회'를 제시하는데, 시민의 보편적 동의가 알고리즘에 반영되는 단계를 추가
  - 시민의 보편적 동의는 사회윤리기준으로 구체화할 수 있으며, 이 윤리기준은 정부가 시민사회와의 긴밀한 협의과정을 통해 정립할 필요가 있음
  - '루프 안의 사회'에서는 권리, 윤리, 법, 규범, 프라이버시, 공정성, 사회계약 등 인간사회에 통용되는 가치에 따른 알고리즘 개발과 평가가 작용하는 구조이며 사회와 기술의 공진화를 추구하는 과정

[참고 4] 싱가포르 정보통신규제청의 위험평가 메트릭스



[그림 설명]

- 싱가포르 정보통신규제청은 한 개인에 대한 기관의 의사결정 결과로 개인에게 해를 끼칠 확률과 심각도를 분류하는 매트릭스를 제안하여 평가에 활용
  - 피해의 정의와 확률과 심각도의 계산은 상황에 따라 다르며 부문마다 다름
  - 환자의 건강 상태에 대한 잘못된 진단과 그로 인한 피해는 의류에 대한 잘못된 제품 추천과 그로 인한 피해와 현격히 다름

- 위험 매트릭스를 바탕으로 조직은 의사 결정에 필요한 수준의 인간 참여를 식별
  - 예를 들어 안전이 가장 중요한 시스템의 경우라면 기관은 사람이 AI시스템을 통제할 수 있도록 해야 하며 때에 따라서는 안전하게 종료 할 수 있도록 해야 함

출처 : Discussion paper on AI and personal data, PDPC, 2018

[참고 5] 편향(Bias)의 5가지 종류

❶ 샘플 편향 : 수집된 데이터가 AI 시스템이 실행될 것으로 예상되는 환경을 대표하거나 정확하게 나타내지 않을 때 발생

(예) CCTV 사례 : 지능형 CCTV의 목표가 주간과 야간에 보안 카메라를 작동시킬 수 있는 모델을 만드는 것이지만 야간 데이터세트만 훈련하는 경우 모델에 이미 샘플 편향이 발생

❷ 제외(배제) 편향 : 데이터를 정제할 때, 기존의 데이터 레이블 및 산출물과 관련이 없다고 생각하여 데이터 세트의 일부 속성(특징)데이터를 제외시킨 결과 발생

(예) 타이타닉호 생존자 예측 : 생존자를 예측하는 유명한 타이타닉 문제에서 생존 여부와는 전혀 무관하다고 생각하여 여행자의 승객 ID 데이터를 무시, 실제 데이터세트에서 제외

❸ 옵서버 편향 (실험자 편향) : 실험자가 기대하는 결과를 보고자 하는 경향에서 발생, 실험자가 의식이나 무의식의 편견을 가지고 실험하는 경우 발생 가능

(예) IQ 유전성에 관한 연구(Cyril Burt) : 심리학자 Cyril Burt는 사회 경제적 지위가 낮은 (예 : 노동 계급 자녀) 가족의 어린이가 사회 경제적 지위가 높은 가족의 어린이보다 지능이 낮을 가능성이 높다고 생각

❹ 편견 편향 : 편견은 외모, 사회적 계급, 지위, 성별 등에 관한 문화적 영향이나 고정관념의 결과로 발생

(예) 직장에서 일 할 사람을 찾는 컴퓨터 비전 프로그램 : 직장에서 일 할 사람을 찾는 모델로. 남성은 코딩하고 여성은 요리하는 수천 개의 훈련 데이터가 제공되었을 때 이 알고리즘은 코딩하는 사람은 남자, 요리하는 사람은 여자로 인식

❺ 측정 편향 : 측정에 사용된 장치에 의해 시스템적으로 값이 왜곡되는 문제가 있는 경우 발생

(예) 밝기를 증가시키는 카메라로 이미지를 촬영 : 측정 편향이 있는 도구는 모델이 작동할 환경을 그대로 복제하지 못하며 학습데이터가 실제데이터를 왜곡되어 편향된 결과를 가져오게 됨

출처 : 5 Types of bias, Salma Ghoneim, Data Science in the world(2019)

참고 자료

한국정보화진흥원, “인공지능시대의 정부(인공지능이 어떻게 정부를 변화시킬 것 인가?)”, IT & Future Strategy, June 26, 2017

한국정보화진흥원, “신뢰가능 AI 구현을 위한 정책 방향(OECD AI 권고안을 중심으로)”, IT & Future Strategy, June 19, 2019

박덕영, “OECD 다국적기업 가이드라인과 기업의 사회적 책임”, KERI Column, June 11, 2016

국가인권위원회, “인권영향평가 및 관리에 관한 지침(HRIAM 가이드)”, December 2014

Ferguson, A. (2014), “Big Data and Predictive Reasonable Suspicion”, ,http://dx.doi.org/10.2139/ssrn.2394683.

Brayne, S., A. Rosenblat and D. Boyd (2015), “Predictive policing, data & civil rights: A new era of policing and justice”, ,Vol. 163/327, http://www.datacivilrights.org/pubs/ 2015-1027/Predictive_Policing.pdf.

Joh, E. (2017), “The undue influence of surveillance technology companies on policing”, ,Vol. 91/101,http://dx.doi.org/10.2139/ssrn.2924620.

Pro Publica, “How We Analyzed the COMPAS Recidivism Algorithm” by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin, May 23, 2016

Office of the Inspector General, “Review of selected Los Angeles Police Department Data-Driven Policing Strategies” by Mark P. Smith, March 12, 2019

LA Times, “LAPD Predictive Policing Tool Raises Racial Bias Concerns” by BY Mark Puente, April 11, 2019

Salma Ghoneim, “5 Types of bias & how to eliminate them in your machine learning project”, Data Science in the world

IFC, “Guide to Human Rights Impact Assessment and Management” (http://www1.ifc.org/), Global Compact, International Business Leaders Forum 2010,

Iyad Rahwan, “Society-in-the-Loop, Programming the Algorithmic Social Contract”, MIT media lab, Aug 13, 2016

DARPA. "Explainable Artificial Intelligence (XAI)". DARPA presentation. Retrieved 17 July 2017.

OECD, “Recommendation of the Council on OECD Legal Instruments Artificial Intelligence”, May 22, 2019

Deloitte Insights, “2018 Global Human Capital Trends”, https://documents.deloitte.com/insights/HCTrends2018

Deloitte University Press, “AI-augmented government : Using cognitive technologies to redesign public sector work”, 2017

Julian Torres Santeli & Sabine Gerdon, “5 challenges for government adoption of AI”, WEF(2019)

Ali Hashmi, “AI Ethics: The Next Big Thing In Government”, WGS(2019)

PDPC Singapore, “Discussion paper on AI and personal data-Fostering responsible development and adoption of AI”, June 5, 2018

NIA

DNA플러스 2019

공공기관 신뢰가능 AI 구현 실용가이드

- OECD 권고안의 적용 -

발　행 : 2019. 12.

발행인 : 문 용 식

발행처 : 한국정보화진흥원 정책본부
미래전략센터

기획 및 문의 : 한국정보화진흥원(NIA) 정책본부 미래전략센터
김동현 수석(053-230-1288, kimdh@nia.or.kr)
장준희 선임(053-230-1298, junhee@nia.or.kr)